Title
Cross-source sensing data fusion for building occupancy prediction with adaptive lasso feature filtering

Authors
Wang, W
Hong, T
Xu, N
et al.

Publication Date
2019-09-01

DOI
10.1016/j.buildenv.2019.106280

Peer reviewed

Cross-source sensing data fusion for building occupancy prediction

with adaptive lasso feature filtering

Wei Wang [a], Tianzhen Hong [b], Ning Xu [a], Xiaodong Xu [a1], Jiayu Chen [c, d], Xiaofang Shan [c]

[a] *School of Architecture, Southeast University, Nanjing, Jiangsu Province, China*
[b] *Building Technology and Urban Systems Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[c] *Department of Architecture and Civil Engineering, City University of Hong Kong, Y6621, AC1, Tat Chee Ave, Kowloon, Hong Kong*
[d] *Architecture and Civil Engineering Research Centre, City University of Hong Kong Shenzhen Research Institute, Shenzhen, Guang Dong 518057, China*

Abstract:

Fusing various sensing data sources is able to improve the accuracy and reliability of building occupancy detection. Efficiently fusing environmental sensors and wireless network signals is seldom studied for its computational and technical challenges. This study aims to propose an integrated model that is able to extract critical data features for environmental and Wi-Fi probe dual sensing sources to promote computational efficiency. The adaptive lasso model was introduced for the feature extraction and reduction process. To validate the proposed model, an onsite experiment was conducted and two occupancy resolutions, real-time and four-level occupancy resolutions, were compared. The results suggested that eight features among all twelve features are most relevant. The mean absolute error of the selected data features is about 2.18 for real-time occupancy and F1_accuracy is about 84.36% for four-level occupancy.

Keywords: data fusion, physic-based model, machine learning, feature selection, occupancy prediction

1*Corresponding author. tel.: +86 (25) 8379-5689. E-mail: xuxiaodong@seu.edu.cn (Xiaodong Xu)

1. Introduction

With HVAC (heating, ventilation, air-conditioning) systems consuming over 40% of building energy use, improving efficient HVAC control is a key issue in building energy saving studies [1,2] Under building operation phase, not only can occupants interact with building to maintain indoor thermal comfort and environment quality, also occupants can passively participate in building load transfer, therefore, the influence of occupancy on buildings' performance increases [3,4]. Occupancy detection and prediction are inspiring researches for efficient HVAC controls and developing building energy efficiency models [5,6]. Previously, occupancy was usually estimated with a single parameter with single sensor, e.g. CO2 [7], lighting [8], PIR [9], Bluetooth [10,11], Wi-Fi [12,13], and so on. However, with development of sensor and information technologies, to improve the accuracy and robustness of occupancy detection and prediction, occupancy estimation with multiply sensors/parameters fusion is a significant trend instead of by a single parameter [14–16]. Among data fusion studies, fusing different types of environment sensing, e.g. temperature, relative humidity, and CO2 concentration parameters, and so on, as well as other types of sensing, e.g. Wi-Fi, motion, lighting, and so on, has attracted increasing attentions.

Such studies proved that the data fusion method can achieve good accuracy of detecting and predicting occupancy, however, a parameter selection within fusing environmental data and Wi-Fi data, in general, should be necessiarly explored to determine the best set of different datasets as well as improve accuracy of occupancy prediction. Masood et al. proposed a filter-wrapper component for a feature selection process with fusing indoor CO2, relative humidity, temperature, and pressure levels in the office space, finding that CO2 feature achieved excellent accuracies of up to 81.67% [17]. However, how to determine the best feature set fusing environmental sensing with Wi-Fi and other building operation datasets, has been not usually explored. Therefore, this paper conducts a data fusion method and paramter selection process while combining building operation and Wi-Fi datasets in occupancy prediction in office buildings. Additionally, an insight into data fusion is provided by occupancy feature selection and extraction with physics-based models during building operation. Finally, this study also deepens the exploration of data-driven occupancy prediction using machine learning algorithm to figure out the best data fusion.

2. Background

2.1 Occupancy studies with single data type

Currently, using single environmental parameter or sensor technology to detect and predict occupancy has been studies in many works. The most popular parameter is indoor CO2 concentration. Wang et al. applied CO2 sensor to monitor concentration and dynamic CO2 concentration physical balance function to predict occupancy count information [18,19]. Díaz and Jiménez proposed an experimental study to assess building occupancy pattern through CO2 concentration and compared it with computer power consumption [20]. Jiang et al. estimated indoor occupancy information through a feature-scale learning machine with measured CO2 concentration dataset [7]. Yang et al. compared the four different occupancy counting methodologies, overhead video, pan-tilt-zoom (PTZ) camera face detection, CO2-based physical model, and CO2-based statistical model, and results showed that the PTZ-camera based face recognition has the most stable and highest accuracy with an R2 of 0.972, followed by the CO2 based statistical model with an R2 of 0.938. Ouf et al. used experimental datasets and Pearson’s correlations to investigate both Wi-Fi connections and CO2 concentration-based approaches for occupancy assessment, suggestting that Wi-Fi counting is more accurate and reliable [21].

As proved, Wi-Fi connections and disconnections can also be utilized as indicators of building energy load variation [22] and occupancy pattern [23]. Balaji proposed a study proving an 83% accuracy of detecting occupancy profiles via Wi-Fi connections [24]. Wi-Fi technology has been applied in occupancy patterns and energy efficiency studies [25,26]. Wang et al. used Wi-Fi discontinuous wireless communication to detect occupancy via event-triggered updating method and achieved the accuracy of at least 77.3% [27]. Wang and Shao conducted one 24-h monitoring over 30 days in library and applied a rule mining approach, finding 26.1% of total energy cost can be saved [28]. Since Wi-Fi signals distribute indoor space like air surrounding it and will be reflected by human body, MIT researchers conducted an experiment to identify occupant and the gesture with Wi-Fi technology through walls during indoor space [29]. Wang et al. also explored the Wi-Fi probe based occupancy study to sense the Wi-Fi signal request and response and achieved over 80% accuracy of occupancy detection [30]. Using Wi-Fi technology to control building lighting as well as occupancy detection, Zou et al. demonstrated the 93.09% and 80.27% of energy saving instead of static scheduling and PIR based lighting control scheme [31].

Also, in some studies, indoor lighting is a kind of parameter to monitor occupancy information by e.g. visible light communication technology [32]. Yang et al. inferred occupancy counting via multiply LED sensing with indoor lighting infrastructure with experiments in a 30 m^2 office area [8]. Park et al. applied LightLearn method that learns the individual occupant behaviors with reinforcement learning algorithm to build occupant centered control based lighting system for energy saving [33]. On the other hand, Manzoor proposed a study for efficient building lighting control by monitoring occupancy with passive RFID technology, which proved 13% of electrical energy savings [34]. Li et al. reported the average accuracy of RFID systems was 88% for stationary occupants and 62% for mobile occupants [35].

2.1 Occupancy studies with data fusion

Besides single parameter or sensing technology, a recent development has been the use of environment sensors for occupancy estimation with environmental parameter array. The relationship between occupancy,multi-environmental parameters, and other sensors has been established and proven to be very useful in occupancy models [36–38]. Pedersen et al. applied an occupancy detection method using air temperature, humidity, CO_2, and VOC, PIR noise sensors. The experiment was conducted in a simple test room and a three-room dorm to detect two occupancy statuses of room, occupied or vacant, resulting in a maximum accuracy of 98% and 78%, respectively in two rooms [39]. Roselyn et al. used thermal sensors and camera to detect occupancy and applied image processing algorithm and sensor signal processing algorithms for energy-efficient control [40]. Jeon proposed Internet-Of-Thing (IoT)-based occupancy detection with fusion of dust ($PM_{2.5}$ and PM_{10}) concentration, humidity, and temperature sensors [41]. Related to Soh's studies [42], occupancy estimation has been studied by considering temperature, RH, CO_2, air pressure. Several algorithms with environmental sensing data has been discussed individually, namely Location Receptive Fields, ANN, k-NN, SVM, CART, extreme learning machine, liner discriminant functions (LDA). Szczurek et al. studied the performances of three environment parameters, temperature, RH, and CO_2 individually and the three sensors array in occupancy determination. The authors also compared k-NN algorithm and LDA when occupancy classification was required, where k-NN was more efficient [16]. To find occupancy in large-scale area, Dong et al. [37] applied one information technology enabled sustainability test-bed (ITEST) for occupancy detection with a wireless ambient-sensing system, a wired carbon dioxide sensing system, and a wired indoor air quality sensing system. The experiment was conducted in a large-scale open office area and it resulted in an average of 73% accuracy in such areas. Based on machine learning techniques, Ryu and Moon developed one occupancy prediction model using CO2, 1st order shifted of difference of CO2, indoor CO2 moving average and rate of change, and indoor and outdoor CO2 ratio as indoor environmental data feature [43]. Two data-driven decision tree and hidden Markov model (HMM) algorithms were proved well suited to detect occupancy. With a fusion of light sensor, Candanedo and Feldheim, also evaluated a method of temperature, humidity and CO2 sensors to predict occupancy with different statistical classification models, LDA, Classification and Regression Trees (CART), and Random Forest (RF). They found about 97% accuracy when using only two of environmental parameter with LDA model in one-day measurement. For example, Zhu et al. estimated office occupancy with environmental sensing via non-iterative local receptive fields in time and frequency domains with a data conclusion of CO2, humidity, temperature, and air pressure. Becerik-Gerber et al. studied a fusion of light, sound, motion, CO2, temperature, relative humidity, PIR, door switch sensors and applied ARMA, Nerual Network, Markov Chain, and Logit Regression to model occupancy profiles [44]. Wang et al. proposed a study of predicting occupancy information through data fusion of environmental sensing and Wi-Fi dataset and applied machine learning techniques to figure out the most accurate set [45]. Chen et al. proposed a novel fusion with Wi-Fi and Bluetooth Lower Energy (BLE) network to collecting building occupancy distribution using different signal distance measurement metrics [46].

The approaches reviewed above employed single/multiply sensing technologies and for occupancy prediction as well as various data-driven algorithms embedded with sensing technologies. To reduce cost, efficiency, and accuracy of occupancy prediction, this study would like to conduct the data feature extraction and parameter selection processing that fuses different sets of multiply parameters within building physic- and machine learning-based models.

3. Methodology

3.1 Dataset feature extraction

Usually, no matter data from Wi-Fi signal and environmental sensing, these are the time series data, which might consist of dataset default and abnormal data point. While pre-processing raw data, the Exponential Moving Average (EMA) filter is applied due to its computational efficiency and causality which are also important in time-series applications. It can be formulated:

$$\acute{x}_k = \frac{n}{n+1}\acute{x}_{k-1} + \left(1 - \frac{n}{n+1}\right)x_k \qquad (1)$$

Where $\acute{x}_k$ and $\acute{x}_{k-1}$ are the EMA filtered value at time step k and k-1, respectively. Once the measured indoor environment data has been filtered, the following approach is applied to detect occupancy.

3.1.1 Features from physical equations

For the feature-based occupancy prediction, feature is a variable which contains the information relevant for object recognition, while in occupancy study, it refers to relevant information for occupancy determination. The basis for choosing appropriate variables for occupancy determination was a well-known fact that properties of indoor air, for example, CO2 concentration, RH, or temperature, have been proven as triggers to stimulate occupancy behaviors to restore or improve comfort conditions [47]. The value and its change of environment parameters should refer to the corresponding occupancy profile since building will response to occupancy behavior and adjust to meet occupant thermal comfort if building is occupied. To figure out the parameters determining occupancy, the common and simplified mass and energy balances functions are analyzed.

For indoor air quality control (it supposes in this study that only

CO2 concentration is considered), assuming the CO_2 only comes from occupant respiration and outdoor air, and CO_2 generation (S) from the occupant is kept constant and the CO_2 concentration (C_o) of outdoor air doesn’t vary by a wide margin. The air supplied to space is assumed to be well-mixed. The time variation of CO_2 concentration levels in one zone can be given based on mass balance equation:

$$V_{room}\frac{d\,C_{¿}(t)}{dt}=V_{sa}C_{sa}+P_{z,i}*S-V_{ra}C_{ra}-V_{oa}C_{ra}+C_{other} \qquad (2)$$

While in the AHU, mass balance of CO2 yields:

$$V_{oa}C_o+V_{ra}C_{ra}=V_{sa}C_{sa} \qquad (3)$$

The$C_{¿}$is the indoor CO_2 concentration and C_{rtn} is the CO_2 concentration at return duct level. Assuming CO2 concentration at return air ducts keeps the same as the CO2 concentration of indoor air at breathing level, we could simply Eq.3 as:

$$V_{room}\frac{d\,C_{¿}(t)}{dt}=V_{oa,z,i}C_0+P_{z,i}*S-V_{oa,z,i}*C_{z,i}+C_{other} \qquad (4)$$

Assume the density of outdoor air are the constant, therefore, the occupant count can be recognized as a function of outdoor air flow rate (m_{oa}) and indoor air CO2 concentration, which can be roughly expressed as:

$$P_z \leftarrow f\left(m_{oa}, m_{oa}*C_{¿}\right) \qquad (5)$$

The similar case of relative humidity can be inferred as CO2 concentration when mass balance is applied. For brevity, this study doesn’t present the derivation of equation. However, the relative humidity of outdoor air usually changes with time rather than keeps constant as CO2 concentration. Therefore, the function between occupant count and outdoor air flow rate and relative humidity can be roughly expressed as:

$$P_z \leftarrow f\left(m_{oa}*RH_{out}, m_{oa}*RH_{¿}\right) \qquad (6)$$

Where RH_{out} and $RH_{¿}$ are the RH of outdoor air and indoor air.

For indoor thermal comfort control, energy balance can be applied that the supplied energy should be equal to the consumed energy, which can be followed by Eq. 7,8, and 9.

$$Q_{supply}=Q_{vent,r}+\sum_{P_z} Gp+\sum_{p_{eq}} G_{eq}+\sum G_{other} \quad (7)$$

$$Q_{supply}=m_s*C_p*\left(T_{¿}-T_s\right) \quad (8)$$

$$Q_{vent,r}=m_{oa}*\left(h_{oa}-h_{¿}\right) \quad (9)$$

Where Q_{supply} is energy supplied, $Q_{vent,r}$ is the energy produced by ventilation, Gp is the energy produced by each occupant. $p_{eq}$and G_{eq} donate the number of equipment and the energy produced by equipment, such as computers, water heaters, lights etc. G_{other} includes the energy produced by other sources, such as infiltration air, adjacent walls, surface and etc. m_s is the supply air flow rate. $T_{¿}$ and T_s are the temperature of indoor air and supply air.h_{oa} and $h_{¿}$ are the entropy of outdoor air and indoor air, which are the function of temperature and RH of air.

Similarly, the function between occupant count and operation parameters can be roughly expressed as:

$$P_z \leftarrow f\left(m_{oa}*T_{out}, m_{oa}*T_{¿}, m_s*T_s, m_s*T_{¿}\right) \quad (10)$$

3.1.2 Features from Wi-Fi connections

In Wi-Fi data sensing, time tag is used to calibrate environmental parameters and Wi-Fi data and the total number and frequencies of Mac addresses in the time window can be the features. The number can be found by counting the valid Mac addresses in one time-spot. The frequency can be represented by the probability that one Mac address will be detected, which can be in details found in [30]. The probabilities are calculated by:

$$f_m^{i-i}=\frac{\sum N_{i-i}}{\sum N_{i-i}+\sum N_{i-i}} \quad f_m^{o-i}=\frac{\sum N_{o-i}}{\sum N_{o-o}+\sum N_{o-i}} \quad (11)$$

Where N_{i-i} is the frequency that occupancy status transited from “in” to “in” and $N_{i-o}$is the frequencies that occupancy status transited from “in” to “out” respectively. Similarly, N_{o-o} and N_{o-i} mean the frequencies that occupancy status transited from “out” to “out” and from “out” to “in” respectively.

One vector can be defined as feature of Wi-Fi data: $\{N_t, f_1^{i-i}, f_1^{o-i}, f_2^{i-i}, f_2^{o-i}, \ldots, f_m^{i-i}, f_m^{o-i}, \ldots, f_N^{i-i}, f_N^{o-i}\}$, this feature contains the number (N_t¿ of Mac addresses in the time-spot (t), and thereinto each Mac address transits from “in” to “in” and “out” to “in”. Similarly, the function between occupant count and operation parameters can be roughly expressed as:

$$P_z \leftarrow f\left(N, f^{i-i}, f^{o-i}\right) \tag{12}$$

Therefore, a parameter pool can be created with union of Eq. 5, 6, 10, and 12 from physical equations based on mass and energy balances in buildings. As reviewed, since some researchers have investigated and concluded the opportunities of using single environmental parameter of indoor air—temperature, relative humidity, CO2—to sense occupancy information, this study, therefore, takes into consideration of those parameters. The dataset feature pool can be roughly expressed as:

$$P_z \leftarrow f \text{¿} \, Wi\text{-}Fi) \tag{13}$$

3.2 Dataset feature selection

For feature-based prediction model, environmental and Wi-Fi data related features are extracted in this study. However, when the parameters are extracted, the feature selection is an important issue in feature driven occupancy estimation. Theoretically, the combination of features we can use is C_{12}^{n} (n=1, 2, 3, …, 12). Generally speaking, accuracy of occupancy estimation can be improved when more multiple features are selected while the computational burden is quite high. On the other side of the coin, extra data collection always means higher cost. In such fusion, it would have substantial benefits and practical implications if an adequately high prediction accuracy could be achieved with as few inputs as possible [48]. In this study, two steps are conducted. The first step is to select features that mostly correlate to occupancy profiles. Secondly, multi selected features will be compared and evaluated in occupancy model for the final feature set.

3.2.1 Feature selection from correlation analysis

In the first step, the best features for each parameter is chosen to reduce feature space to a more manageable number. The historical data match between features and occupancy profile will be used to reveal the relationships from features to occupancy. Such problems usually be solved with least squares method, which is usually applied in regression, or stepwise regression, which is usually used in feature selection for prediction. This study takes the Adaptive-Lasso model to reveal the correlation between different data features and occupancy profile and select best data features. Adaptive-Lassois one of the methods widely used in parameter estimation and variable selection []. The definition is:

$$\hat{\beta}^{¿(n)} = \underset{\beta}{argmin} \left| y - \sum_{j=1}^{p} x_j \beta_j \right|^2 + \alpha_n \sum_{j=1}^{p} \hat{\omega} \vee \beta_j \vee ¿¿ \qquad (14)$$

Where,

With the feature spaces reduce, the best features from the first step are combined as multi feature set to finally evaluate their performance. All possible combinations of the elements of this set were examined with the ANN model

3.2.2 Feature selection from occupancy prediction

To avoid inaccurate predictions due to the magnitude of the data, all input variables are normalized to [-1,1] according to the Eq.15.

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (15)$$

A Backpropagation artificial neural network (BP-ANN) algorithm is a computational method used to calculate the weights and minimize system error, which can be used for regression analysis, including time series prediction and modeling, and classification analysis, including pattern recognition and sequential decision making. It employs the gradient descent optimization to adjust the weight of neurons until the gradients of the loss functions are minimized.

The main equations of BP-ANN are summarized as Eq. 16 and 17. The x_n represents any selected occupancy feature and n is the dimension of features. The m and l are the number of the hidden layer and output layer neurons. The v_j donates the weight vector of the jth neuron of the hidden layer, and w_k donates the weight vector of the kth neuron of the output layer. The length of the input layer is determined by the available data sources, while the size of the hidden layer (m) is manually selected. The size of the output layer (l) usually equals the number of expected output elements.

$$h_j = f\left(\sum_{i=0}^{n} v_{ij} x_i\right), j = 1, 2, \ldots, m \tag{16}$$

$$y_k = f\left(\sum_{j=0}^{m} w_{jk} h_j\right), k = 1, 2, 3, \ldots, l \tag{17}$$

The sigmoid function is usually selected as the transfer/activation function, as shown in Eq. 18.

$$f(x) = \frac{1}{1+e^{-x}} \tag{18}$$

4. Experiment and Validation

4.1 Experiment Setup

The experiment test bed is a graduate student office located inside an institutional building. The office has an area of about 200 m^2 and 25 long-term residents during the experiment period. Figure 1 shows the space layout and equipment setup of the test bed. The office has two entrances but no window. Wi-Fi probes recorded the connection requests and responses of all wireless devices within the space. TA465-X (environmental sensors produced by TSI Company) were utilized to monitor and record the air temperature, relative humidity, and CO2 concentration. Air flow meters were installed near outdoor air inlets to monitor the air supply rate of the ventilation system. Ground truth is acquired by two overhead cameras installed to record the entrance and exit events of occupants. Since sample time of sensors is different, we need to firstly obtain the entrance and exit of doors from videos at any time and easily calculate the number of occupants at same sample time as Wi-Fi probe and CO2 concentration. The measurement duration is from 09 Sep 2017 to 23 Sep 2017. Table 1 shows the specifications of the installed sensors.

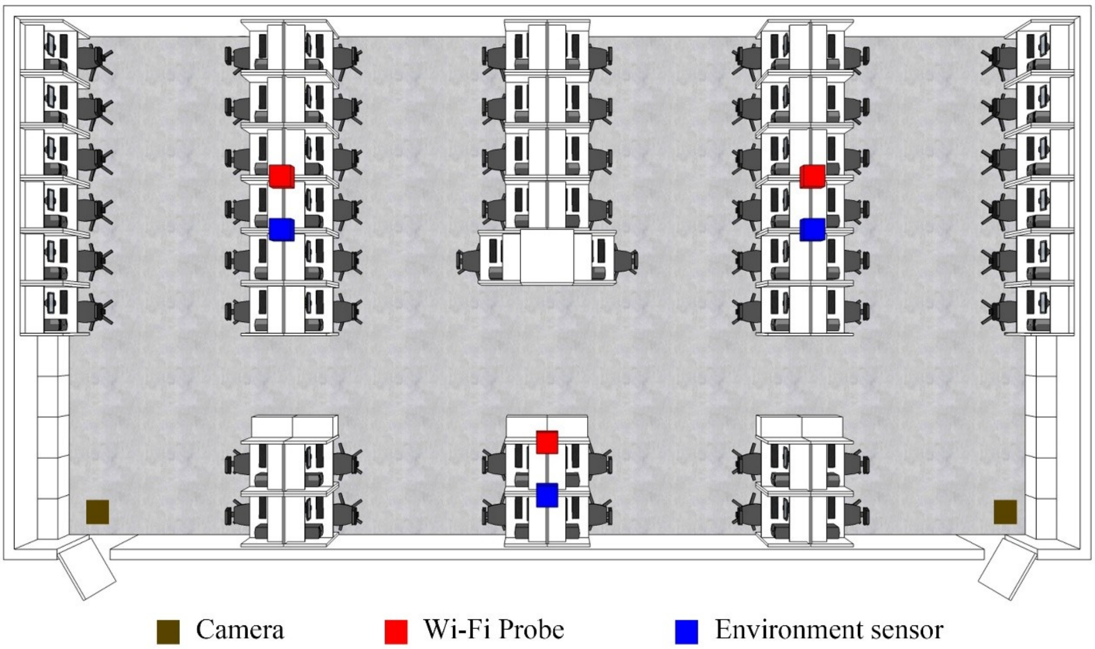

Figure 1. Space layout and equipment setup

Table 1. Sensors used in the experiment.

Sensor	Camera	Wi-Fi Probe	Environment Sensors			
			CO2 Sensors	Temperature Sensors	Humidity Sensors	Other Sensors
Cost (USD)	45	30	400			
Recorded Variables	Time, Actual occupancy	Time, MAC address, RSSIs	Time, Temperature, Relative humidity, CO2, Air flow rate, Air pressure, CO			

Data Storage	Online	Online	Local		
Measurement timestep		30s	1min	1min	1min
Range			0 – 5k ppm	14 - 140 °F -10 – 60 ℃	0 to 95%
Accuracy			±3% or ±50 ppm	±0.5°F (±0.3℃)	< 3%
Resolution			1 ppm	0.1°F (0.1℃)	0.10%

4.2 Ground truth acquisition

Overhead cameras were installed close to the doors to capture the actual number of occupants in each office room. In addition, two cameras in room A were synchronized with internet time. Because the sampling data of the Wi-Fi probe device was one minute, the video analysis obtained the number of entrances and exits through each door at the same sampling frequency. The number of occupants were manually counted based on the recorded video for each minute.

4.3 Model configuration and assessment

After obtaining the data from the sensors above, the pre-processing is conducted to the raw data. The interval for the TA 465-X sensors is 30 seconds while the recording interval is 1 minute, where the data recorded every minute is averaged by the data every 30 seconds. Originally, the length of samples in one day is 1440. As data comes from three sensors inside room, the final results from measurement should be averaged by three sensors. The Wi-Fi data is the Mac address of user's device, recorded by Wi-Fi probe sensors every 30 second from the Wi-Fi signal request and response between device and access point, while the final Wi-Fi probe data should be merged by three Wi-Fi probes.

While setting the occupancy prediction, as energy optimization and control methods normally do not require the exact number and environment parameters usually respond slowly to control methods [14,49], therefore, this study would like to search the set of data to predict occupancy profile and the occupancy profile contains four levels, including zero, low, medium, and high. It can be expected that with four levels of occupancy, four significantly different range of thermal loads can be identified. Higher levels response to a higher load, for which greater energy is required to maintain the temperature set point. Therefore, this type of occupancy can make HVAC control more simplified and efficient based on four-level demands.

Table 2. The threshold setting for categorical occupancy levels.

Occupancy level	**Number of people**
Zero (0)	0
Low (25%)	1-6
Medium (50%)	7-14
High (75%)	15-20

Therefore, this study would like to compare the predictions of two occupancy types—real-time occupancy and four-level occupancy—to check the performance of different occupancy feature sets on different resolution of occupancy.

5. Results and assessments

5.1 Results of feature selection

Fig. 2 and 3 present the feature selections of dataset for real-time and four-level occupancy prediction. As resulted from AdaptiveLasso model, this study firstly filters out the features highly correlated to actual occupancy datasets. It is interesting to find that the features for two types of occupancy are totally the same, which make it consistent for different occupancy prediction using the same datasets. While inferred from results, the outdoor air flow rate (M_{oa}), and three related features are filters since they made little contributions to occupancy according to AdaptiveLasso model. Additionally, the indoor air relative humidity ($RH_{¿}$) and a set of M_{oa} and outdoor air temperature (T_{out}¿ show the negative correlation. Comparing Fig. 2 and 3, it is interesting to find that feature selection results are consistent for both real-time and four-level occupancy prediction, however, the correlation results are a little different. For example, the set of M_{oa} and CO2 concentration ($C_{¿}$) correlates more highly with real-time occupancy than only $C_{¿}$ does while on the contrary for four-level occupancy. Still, it can conclude that *Wi-Fi* and $C_{¿}$ share the highest correlation with both occupancy levels.

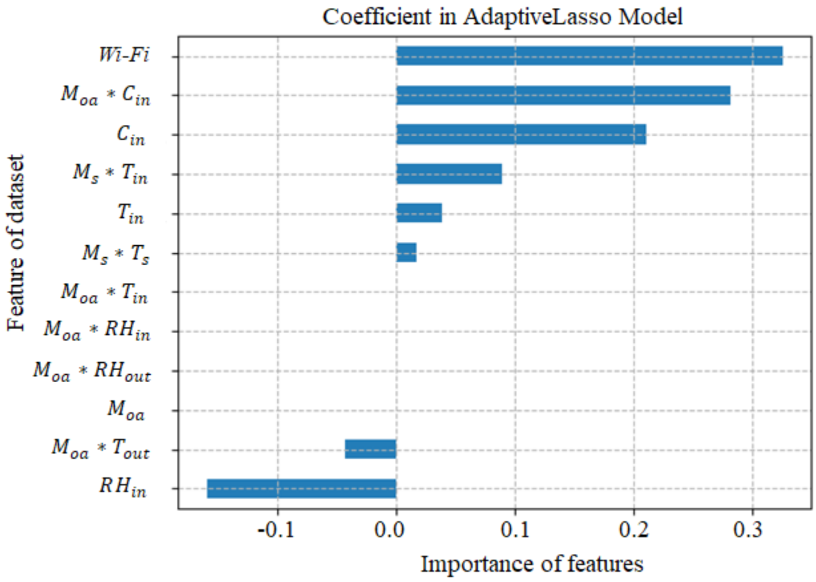

Fig. 2. Feature selection for real-time occupancy prediction using AdaptiveLasso model.

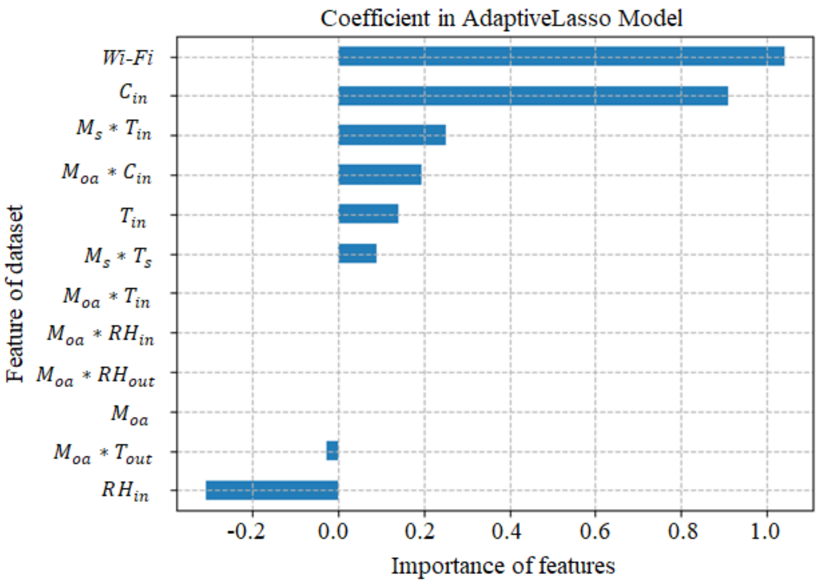

Fig. 3. Feature selection result for four-level occupancy prediction using AdaptiveLasso model.

5.2 Results of occupancy prediction

Inferred from Fig. 2 and 3, eight types of features can be extracted for occupancy prediction, however, among eight features, it can result in 255 sets of feature combination, which can be calculated by $\sum C_8^n$, where (n=1,2,3,...,8). To figure out the best set of feature, this subsection presents the results of using machine learning to investigate prediction accuracy using ANN algorithm for both real-time and four-level occupancy.

5.2.1 Results of real-time occupancy prediction

Fig. 4 presents the results of MAE assessment with different feature sets from only one feature set to seven feature sets. The MAE result is about 89.21% when using all features (C_8^8¿. The results show that the accuracies can be improved as the number of features increases. The best accuracy using only one feature is around 86% using CO2 concentration or Wi-Fi dataset and it is around 88.3% using two features $\left(C_8^2\right)$. Additionally, it sees that when one more feature $\left(C_8^3\right)$ is involved, the best accuracy can be 89.2% as well as the sets of C_8^5 and C_8^6. On one hand, according to the cumulative curve, the results show that usually over 80% of sets can achieve 87.4%. The error distribution tends to higher accuracies along with the increasing of features, which means the possibility of achieving high accuracies can increase with more parameters. The best accuracy for the set of seven features $\left(C_8^7\right)$ is 89.31%. On the other hand, it is also interesting to find that the best accuracy of occupancy prediction can't benefit from the increasing of the number of features since the best accuracy can't be higher than 89.31% in this case study. However, increasing the number of features usually leads to the increasing of sensor cost. Therefore, the number of feature is a trade-off between accuracy and cost. Fig. 5 presents the results of MAE distribution with different parameters for real-time occupancy. The results show that all parameters can contribute to the best accuracy (89.21%), however they also show that the CO2 concentration $\left(C_{¿}\right)$ and Wi-Fi have the best contributions to the prediction accuracy since in the results, the low accuracy of those feature sets assigned with those two parameters will be higher than other parameters and over 95% of feature set can achieve the accuracies of around 86.5% . The results also are consistent to it in Fig. 2.

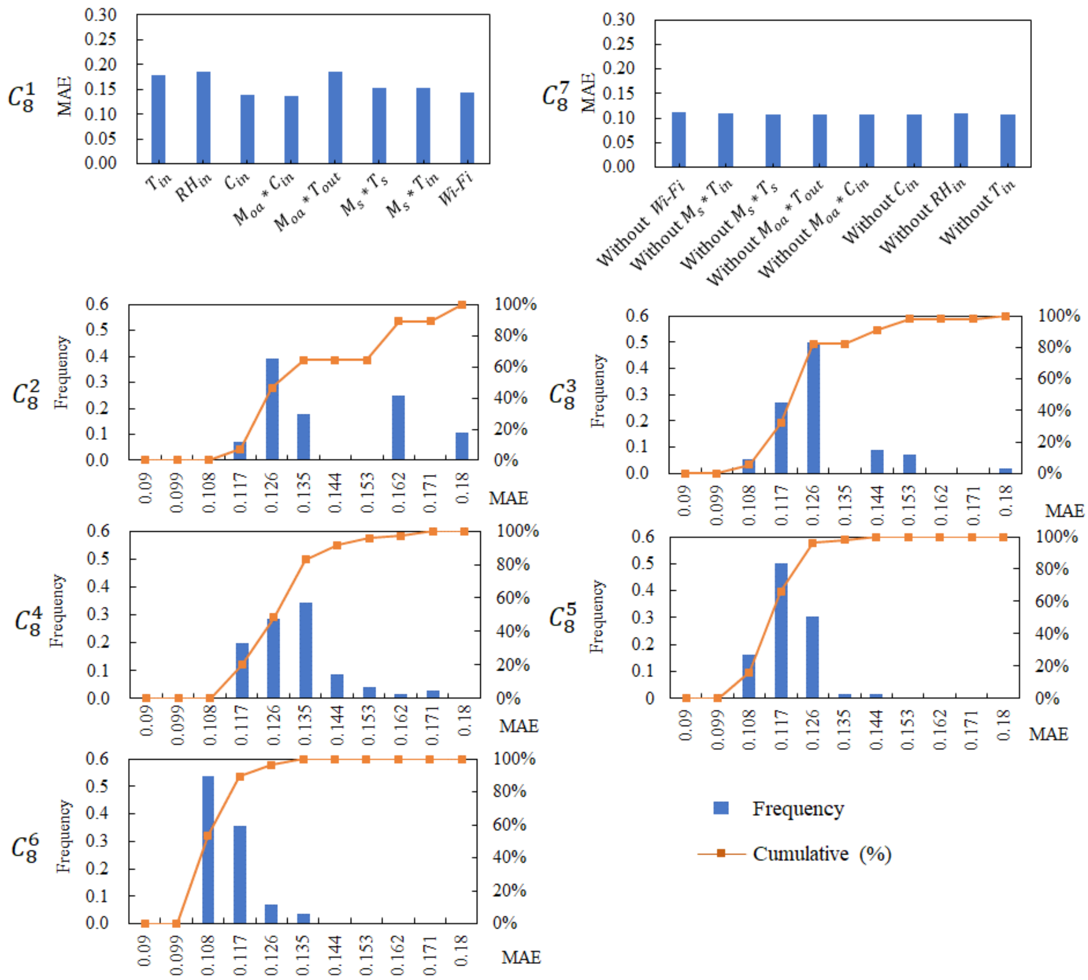

Fig. 4. The results of MAE assessment with different feature sets for real-time occupancy.

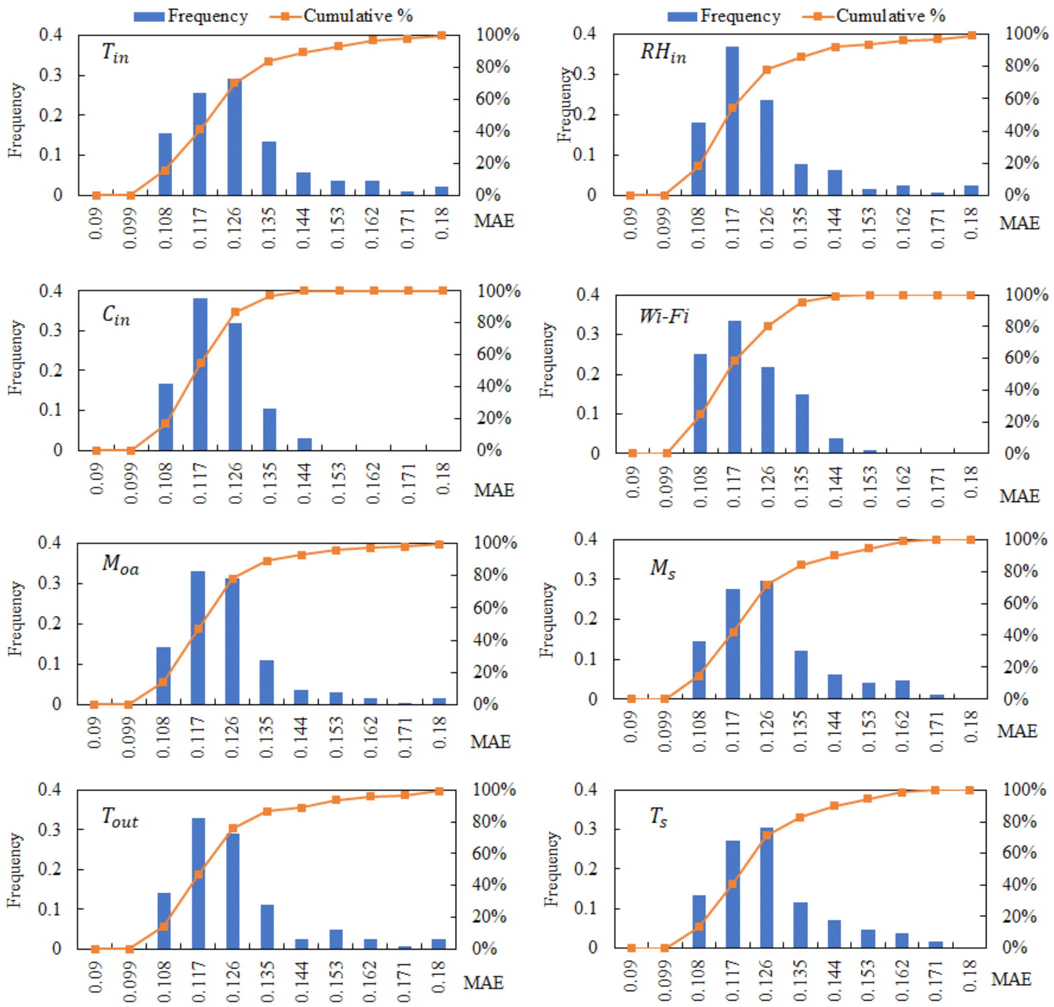

Fig. 5. The results of error distribution with different parameters for real-time occupancy.

5.2.2 Results of four-level occupancy prediction

Fig. 6 presents the results of F1_score assessment with different feature sets from only one feature set to seven feature sets. The F1_score result is about 83.23% when using all features (C_8^8¿. Fig. 7 presents the results of F1_score distribution with different parameters for real-time occupancy. On one hand, the results of four-level occupancy prediction are quite similar to real-time occupancy prediction that the increasing the number of feature can improve the prediction accuracies, however, the best accuracy of occupancy prediction can't benefit from the increasing of the number of features the best, either. On the other hand, it is also interesting that the accuracies using the same eight features are lower in predicting four-level occupancy than real-time occupancy, which might infer that to divide the occupancy information to some levels can enlarge the uncertainty and stochastic behavior of occupancy, especially around boundary of occupancy level.

As seen in Fig. 6, The best accuracy using only one feature is around 81% using CO2 concentration, which indicated that CO2 concentration is a good indicator when applying only one parameter with machine learning techniques in occupancy prediction. The best accuracy is around 84% using two features (C_8^2). Additionally, it sees that when one more feature (C_8^3) is involved, the best accuracy can be 86% as well as the sets of C_8^5, C_8^5 and C_8^6. With involving more features, the proportion of best accuracy (86%) increases. However, it is only 83.74% for seven features (C_8^7) close to seven features (C_8^8). According to Fig. 7, Wi-Fi feature shares the biggest proportion of achieving the best accuracy of 86% while CO2 concentration feature achieved the prediction accuracy of at least about 80%, which is the good indictor in this study.

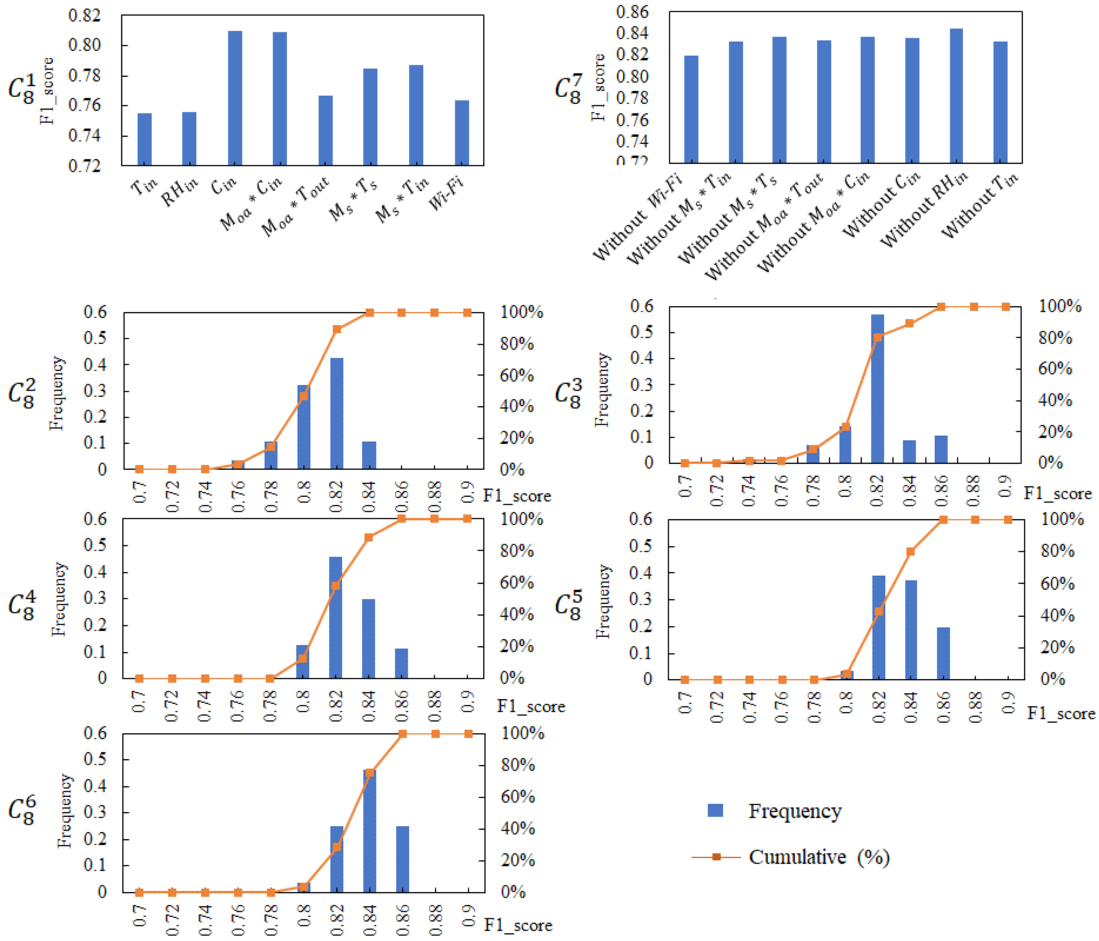

Fig. 6. The results of F1_score assessment with different feature sets for four-level occupancy.

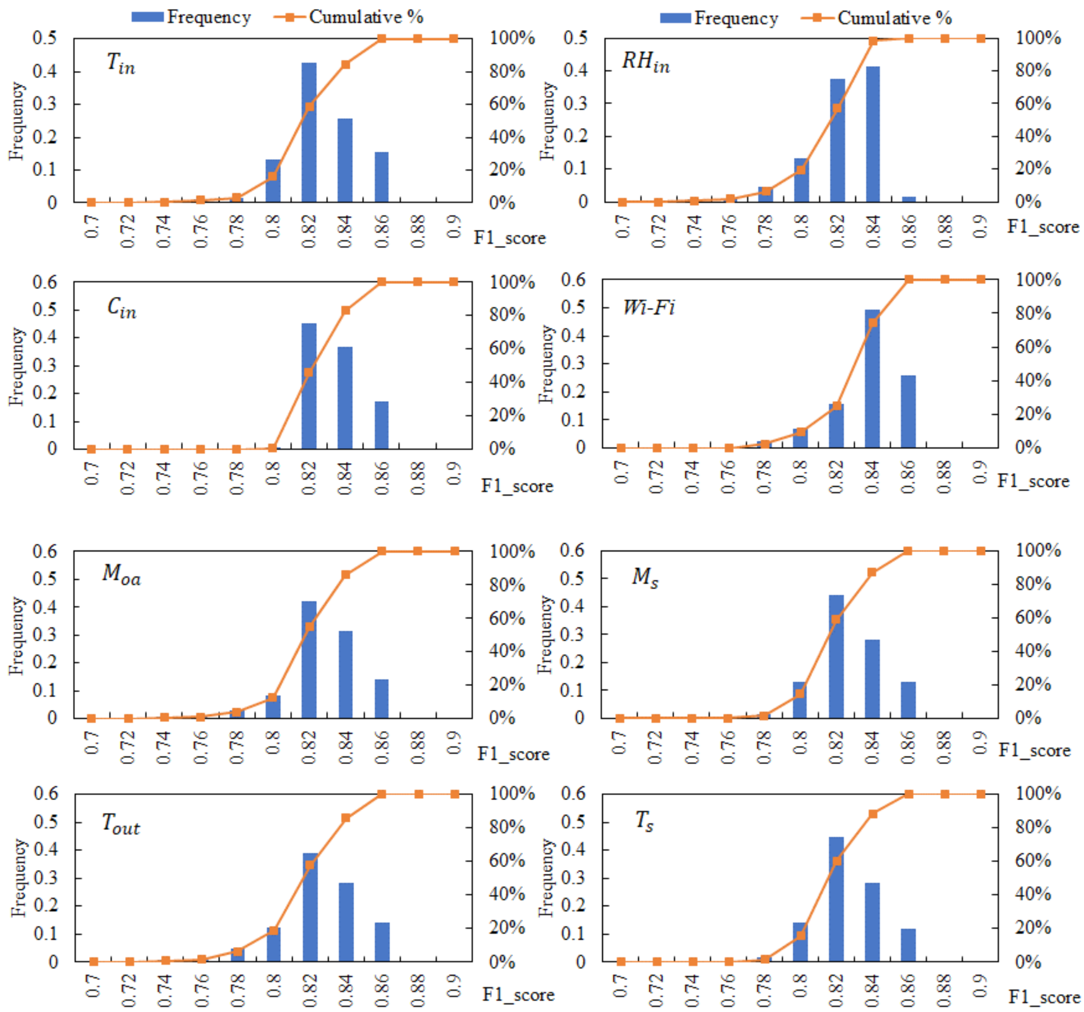

Fig. 7. The results of F1_score distribution with different parameters for four-level occupancy.

5.2.3 Results of parameter selections

After parameter selection through AdaptiveLasso and ANN models, this subsection gives the final results of the best set of feature selection according to the accuracy results. As inferred in Fig. 4 and 5, the most suitable number of feature selection for real-time occupancy prediction is three as well as for four-level occupancy prediction. The next step is to figure out three features achieving around the best accuracy of 89% and 84%, respectively for real-time and four-level, however, it is easily found that there are several choices, 30 for real-time occupancy and 9 for four-level occupancy. For brevity, this study concluded the final results in the Table 3 that include around 84% of F1_accuracy for real-time and

11% of MAE, respectively. It finds that three parameters, indoor air temperature, CO2 concentration, and Wi-Fi dataset, can achieve good prediction results for both real-time and four-level occupancy types, and it involves only three parameters in terms of sensor cost.

Table 3. The final result for parameter selection for real-time and four-level occupancy prediction.

Parameter set selection	Real-time occupancy (MAE)	Four-level occupancy (F1_score)
$T_{¿}+M_{oa}*C_{¿}+¿$ *Wi-Fi*	11.51%	84.71%
$T_{¿}+C_{¿}+¿$ *Wi-Fi*	11.46%	84.36%
$C_{¿}+M_{oa}*T_{out}+¿$*Wi-Fi*	10.97%	84.36%
$C_{¿}+M_{s}*T_{s}+¿$ *Wi-Fi*	10.89%	84.22%
$C_{¿}+M_{s}*T_{¿}+¿$*Wi-Fi*	10.64%	84.13%
$C_{¿}+M_{oa}*C_{¿}+¿$*Wi-Fi*	11.31%	84.10%
$M_{oa}*C_{¿}+M_{s}*T_{¿}+¿$*Wi-Fi*	10.68%	83.87%
$M_{oa}*C_{¿}+M_{s}*T_{s}+¿$*Wi-Fi*	10.88%	83.86%
$M_{oa}*C_{¿}+M_{oa}*T_{out}+¿$ *Wi-Fi*	10.96%	83.72%

[illegible]$T_{¿}+C_{¿}+¿$ *Wi-Fi* [illegible] **real-time** [illegible] **four-level** [illegible] **prediction** [illegible]C_3^1[illegible]C_3^3[illegible]C_3^3[illegible] **1.** [illegible]**2.** [illegible]

6. Discussion

This study investigated the data fusion research for building occupancy prediction to figure out the better dataset combination and more suitable parameters through building operation and Wi-Fi datasets. Two kinds of occupancy information were selected in this study, real-time and four-level occupancy. The parameter selection process was extracted from the building operation process and the indoor mass or energy balance theory as the physics-based models. Usually such model can also infer or predict much accurate occupancy once all parameters can be measured, which is also famously applied as inverse modeling approach [50]. For example, as in Eq. 4, once the CO2 concentration from other sources (e.g. air infiltration) can be accurately measured, we can apply the CO2 mass balance from the sensor data to infer occupancy, which, however, can be a difficult work. The parameter selection framework in this study consists of physical-based models and machine learning techniques to make up for such defects and the framework also provides an insightful reference for data fusion works in occupancy studies. The best accuracies for real-time and four-level occupancy levels are about 90% and 86%, respectively. Inferred in the results, occupancy prediction accuracies can be improved as parameter inputs increase no matter real-time and four-level occupancy levels. However, results reveal that more than four parameters can't improve accuracies a lot and sensor cost is also important issue, even this study didn't make a tradeoff between accuracy and sensor cost, it can usually be considered that increasing the number of parameters will definitely increase the cost. Therefore, on one hand, it recommends using less or cheap sensors for inferring occupancy, on the other hand, it can reduce the sensor cost by sacrificing accuracies since the results in this study show that the best accuracies using one parameter and two parameters for real-time occupancy predictions can reach 86% and 88%, respectively, and 81% and 84% for four-level occupancy predictions, respectively.

On one hand, in this study, results indicate that the combination of temperature, CO2 concentration, and Wi-Fi datasets can have the best accuracies both for real-time and four-level occupancy predictions. As it can see, those three parameters are very common ones in building operation. More significantly, indoor air temperature responses to building cooling/heating systems and CO2 concentration responses to building outdoor air control systems, accordingly, two parameters are usually monitored in building systems. As Wi-Fi signal is almost now available in all buildings, those three parameters are very easily accessed, which benefits a lot for monitoring and predicting occupancy. On the other hand, in terms of control efficiency and robustness, some researchers would like to simplify building control systems using different-level occupancy instead of real-time occupancy as reviewed, since which, therefore, this study investigated the occupancy divided in four different levels. Different occupancy levels refer to different occupant' demand, thereby, this study can benefit those which would like to apply different kinds of occupancy through temperature, CO2 concentration, and Wi-Fi datasets for their building control accuracies.

However, this study yields some limitations. Firstly, as stated in some studies, the occupant impact on indoor air is contained in values of these parameters, but may also retrieved from their changes [51], therefore, it is also an interesting and inspiring work to consider the values of selected parameter changes, which is ignored in this study. Secondly, such study relied a lot on the experiment implementation, e.g. the accuracy, scale, parameter types of experiment monitoring. Future work can bring in more kinds of sensor types and experiment spaces for a larger scale (e.g. floor and building levels) and type group (e.g. lighting and PIR) of occupancy sensing. Furthermore, this study used adaptive-lasso and ANN method as first and second steps to find the best set of data in predicting occupancy. For brevity, this study did not investigate impact of different kinds of algorithms for different sets of data on prediction accuracies, which are interesting future works.

7. Conclusions

Data fusion technology with multiply sensors has attracted more and more attentions in occupancy studies. This study proposed a data fusion study to integrate building physic-based, AdaptiveLasso, machine learning-based models for occupancy feature selection. This study defined two occupancy levels, real-time and four-level occupancy, and conducted one experiment to validate test occupancy feature selection process. In the results, total 12 features were selected from physic-based models and Wi-Fi datasets. Then, AdpativeLasso model figured out eight correlated features and machine learning finally proved three features. The indoor air temperature, CO2 concentration, and Wi-Fi dataset can be fused as the best occupancy feature set with the mean absolute error of about 11.46% for real-time occupancy and F1_accuracy of about 84.36% for four-level occupancy.

This study can contribute to data fusion studies by integrating physical- and machine learning-based models in feature selection for occupancy prediction. Fusing different sensor technologies and data sources for building occupancy prediction can be more efficient and low-cost. In the future, it could be significant using indoor air temperature, CO2 concentration, and Wi-Fi to sense occupancy, in turn to improve building HVAC systems. Also, how to apply such data fusion studies to improve building energy efficiency could be an inspiring work as occupancy prediction accuracy is improved.

Acknowledgement

The work described in this study was sponsored by the projects of the National Natural Science Foundation of China (NSFC#51678127), the National Scientific and Technological Support during the 12th Five-Year Plan Period (No.: 2013BAJ10B13), and Beijing Advanced Innovation Center for Future Urban Design (UDC# 016010100). Any opinions, findings, conclusions, or recommendations expressed in this study are those of the authors and do not necessarily reflect the views of the National Scientific and Technological Support committee and NSFC.

References

[1] D. Yan, T. Hong, B. Dong, A. Mahdavi, S. D'Oca, I. Gaetani, X. Feng, IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings, Energy Build. 156 (2017) 258–270. doi:10.1016/J.ENBUILD.2017.09.084.

[2] A. Wagner, L. O'Brien, Occupant behaviour-centric building design and operation EBC Annex 79 (Draft), (2018).

[3] T. Hong, H.-W. Lin, Occupant Behavior: Impact on Energy Use

of Private Offices, in: Asim 2012, 1st Asia Conf. Int. Build. Perform. Simul. Assoc., 2013.
[4] I. Gaetani, P.-J. Hoes, J.L.M. Hensen, Occupant behavior in building energy simulation: Towards a fit-for-purpose modeling strategy, Energy Build. 121 (2016) 188–204. doi:10.1016/j.enbuild.2016.03.038.
[5] W. Wang, T. Hong, N. Li, R.Q. Wang, J. Chen, Linking energy-cyber-physical systems with occupancy prediction and interpretation through WiFi probe-based ensemble classification, Appl. Energy. 236 (2019) 55–69. doi:10.1016/J.APENERGY.2018.11.079.
[6] C.D. Korkas, S. Baldi, I. Michailidis, E.B. Kosmatopoulos, Occupancy-based demand response and thermal comfort optimization in microgrids with renewable energy sources and energy storage, Appl. Energy. 163 (2016) 93–104.
[7] C. Jiang, M.K. Masood, Y.C. Soh, H. Li, Indoor occupancy estimation from carbon dioxide concentration, Energy Build. 131 (2016) 132–141. doi:10.1016/j.enbuild.2016.09.002.
[8] Y. Yang, J. Luo, J. Hao, S.J. Pan, Counting via LED sensing: Inferring occupancy using lighting infrastructure, Pervasive Mob. Comput. 45 (2018) 35–54. doi:10.1016/j.pmcj.2018.01.003.
[9] R.H. Dodier, G.P. Henze, D.K. Tiller, X. Guo, Building occupancy detection through sensor belief networks, Energy Build. 38 (2006) 1033–1043. doi:10.1016/j.enbuild.2005.12.001.
[10] W. Wang, J. Chen, T. Hong, Modeling occupancy distribution in large spaces with multi-feature classification algorithm, Build. Environ. 137 (2018). doi:10.1016/j.buildenv.2018.04.002.
[11] T.M. Ng, From “Where I am” to “Here I am”: Accuracy study on location-based services with iBeacon technology, HKIE Trans. 22 (2015) 23–31. doi:10.1080/1023697X.2015.1009411.
[12] W. Wang, J. Chen, X. Song, Modeling and predicting occupancy profile in office space with a Wi-Fi probe-based Dynamic Markov Time-Window Inference approach, Build. Environ. 124 (2017). doi:10.1016/j.buildenv.2017.08.003.
[13] H. Li, E.C.L. Chan, X. Guo, J. Xiao, K. Wu, L.M. Ni, Wi-Counter: Smartphone-Based People Counter Using Crowdsourced Wi-Fi Signal Data, IEEE Trans. Human-Machine Syst. 45 (2015) 442–452. doi:10.1109/THMS.2015.2401391.
[14] Q. Zhu, Z. Chen, M.K. Masood, Y.C. Soh, Occupancy estimation with environmental sensing via non-iterative LRF feature learning in time and frequency domains, Energy Build. 141 (2017) 125–133. doi:10.1016/J.ENBUILD.2017.01.057.
[15] M.O. Z. Yang, N. Li, B. Becerik-Gerber, A multi-sensor based occupancy estimation model for supporting demand driven HVAC operations, Proc. 2012 Symp. Simul. Archit. AndUrban Des. Soc. Comput. Simul. Int. San Diego, CA,USA. (2012) 49–56.
[16] A. Szczurek, M. Maciejewska, T. Pietrucha, Occupancy

determination based on time series of CO2 concentration, temperature and relative humidity, Energy Build. 147 (2017) 142–154. doi:10.1016/J.ENBUILD.2017.04.080.
[17] M.K. Masood, C. Jiang, Y.C. Soh, A novel feature selection famework with Hybrid Feature-Scaled Extreme Learning Machine (HFS-ELM) for indoor occupancy estimation, Energy Build. 158 (2018) 1139–1151.
[18] S. Wang, Dynamic simulation of building VAV air-conditioning system and evaluation of EMCS on-line control strategies, Build. Environ. 34 (1999) 681–705. doi:10.1016/S0360-1323(98)00052-3.
[19] S. Wang, J. Burnett, H. Chong, Experimental Validation of CO2-Based Occupancy Detection for Demand-Controlled Ventilation, Indoor Built Environ. 8 (1999) 377–391. doi:10.1177/1420326X9900800605.
[20] J.A. Díaz, M.J. Jiménez, Experimental assessment of room occupancy patterns in an office building. Comparison of different approaches based on CO 2 concentrations and computer power consumption, Appl. Energy. 199 (2017) 121–141. doi:10.1016/j.apenergy.2017.04.082.
[21] M.M. Ouf, M.H. Issa, A. Azzouz, A.-M. Sadick, Effectiveness of using WiFi technologies to detect and predict building occupancy, Sustain. Build. 2 (2017) 7. doi:10.1051/sbuild/2017005.
[22] J. Chen, C. Ahn, Assessing occupants' energy load variation through existing wireless network infrastructure in commercial and educational buildings, Energy Build. 82 (2014) 540–549. doi:10.1016/j.enbuild.2014.07.053.
[23] C. Martani, D. Lee, P. Robinson, R. Britter, C. Ratti, ENERNET: Studying the dynamic relationship between building occupancy and energy consumption, Energy Build. 47 (2012) 584–591. doi:10.1016/j.enbuild.2011.12.037.
[24] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, Y. Agarwal, Sentinel: Occupancy Based HVAC Actuation using Existing WiFi Infrastructure within Commercial Buildings, in: Conf. Proc. 11th ACM Conf. Embed. Networked Sens. Syst., 2013. doi:10.1145/2517351.2517370.
[25] I. Bisio, F. Lavagetto, M. Marchese, A. Sciarrone, Smart probabilistic fingerprinting for WiFi-based indoor positioning with mobile devices, Pervasive Mob. Comput. 31 (2016) 107–123. doi:10.1016/j.pmcj.2016.02.001.
[26] R.S. Campos, L. Lovisolo, M.L.R. de Campos, Wi-Fi multi-floor indoor positioning considering architectural aspects and controlled computational complexity, Expert Syst. Appl. 41 (2014) 6211–6223. doi:10.1016/j.eswa.2014.04.011.
[27] J. Wang, N.C.F. Tse, J.Y.C. Chan, Wi-Fi based occupancy detection in a complex indoor space under discontinuous wireless communication: A robust filtering based on event-triggered updating, Build. Environ. 151 (2019) 228–239.

doi:10.1016/J.BUILDENV.2019.01.043.
[28] Y. Wang, L. Shao, Understanding occupancy pattern and improving building energy efficiency through Wi-Fi based indoor positioning, Build. Environ. 114 (2017) 106–117. doi:10.1016/j.buildenv.2016.12.015.
[29] F. Adib, D. Katabi, Wi-Vi: See Through Walls with Wi-Fi Signals, (2013). http://people.csail.mit.edu/fadel/wivi/index.html.
[30] W. Wang, J. Chen, X. Song, Modeling and predicting occupancy profile in office space with a Wi-Fi probe-based Dynamic Markov Time-Window Inference approach, Build. Environ. 124 (2017) 130–142. doi:10.1016/J.BUILDENV.2017.08.003.
[31] H. Zou, Y. Zhou, H. Jiang, S.-C.C. Chien, L. Xie, C.J. Spanos, WinLight: A WiFi-based occupancy-driven lighting control system for smart building, Energy Build. 158 (2018) 924–938. doi:10.1016/j.enbuild.2017.09.001.
[32] M. Liu, K. Qiu, F. Che, S. Li, B. Hussain, L. Wu, C. Patrick Yue, Towards indoor localization using Visible Light Communication for consumer electronic devices, in: 2014 IEEE/RSJ Int. Conf. Intell. Robot. Syst., IEEE, 2014: pp. 143–148. doi:10.1109/IROS.2014.6942553.
[33] J.Y. Park, T. Dougherty, H. Fritz, Z. Nagy, LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning, Build. Environ. 147 (2018) 397–414. doi:10.1016/j.buildenv.2018.10.028.
[34] F. Manzoor, D. Linton, M. Loughlin, Occupancy Monitoring Using Passive RFID Technology for Efficient Building Lighting Control, in: 2012 Fourth Int. EURASIP Work. RFID Technol., IEEE, 2012: pp. 83–88. doi:10.1109/RFID.2012.10.
[35] N. Li, G. Calis, B. Becerik-Gerber, Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations, Autom. Constr. 24 (2012) 89–99. doi:10.1016/j.autcon.2012.02.013.
[36] L.M. Candanedo, V. Feldheim, Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models, Energy Build. 112 (2016) 28–39. doi:10.1016/j.enbuild.2015.11.071.
[37] B. Dong, B. Andrews, K.P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, D. Benitez, An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network, Energy Build. 42 (2010) 1038–1046. http://www.sciencedirect.com/science/article/pii/S037877881000023X (accessed May 9, 2017).
[38] Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz, A systematic approach to occupancy modeling in ambient sensor-rich buildings, Simulation. 90 (2014) 960–977. doi:10.1177/0037549713489918.
[39] T.H. Pedersen, K.U. Nielsen, S. Petersen, Method for room occupancy detection based on trajectory of indoor climate

sensor data, Build. Environ. 115 (2017) 147–156. doi:10.1016/J.BUILDENV.2017.01.023.
[40] J.P. Roselyn, R.A. Uthra, A. Raj, D. Devaraj, P. Bharadwaj, S.V.D. Krishna Kaki, Development and implementation of novel sensor fusion algorithm for occupancy detection and automation in energy efficient buildings, Sustain. Cities Soc. 44 (2019) 85–98. doi:10.1016/j.scs.2018.09.031.
[41] Y. Jeon, C. Cho, J. Seo, K. Kwon, H. Park, S. Oh, I. Chung, IoT-based occupancy detection system in indoor residential environments, Build. Environ. 132 (2018) 181–204. doi:10.1016/j.buildenv.2018.01.043.
[42] Z. Gu, Z. Chen, Y. Zhang, Y. Zhu, M. Lu, A. Chen, Reducing fingerprint collection for indoor localization, Comput. Commun. 83 (2016) 56–63. doi:10.1016/j.comcom.2015.09.022.
[43] S.H. Ryu, H.J. Moon, Development of an occupancy prediction model using indoor environmental data based on machine learning techniques, Build. Environ. 107 (2016) 1–9. doi:10.1016/J.BUILDENV.2016.06.039.
[44] Z. Yang, B. Becerik-Gerber, Modeling personalized occupancy profiles for representing long term patterns by using ambient context, Build. Environ. 78 (2014) 23–35. doi:10.1016/j.buildenv.2014.04.003.
[45] W. Wang, J. Chen, T. Hong, Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings, Autom. Constr. 94 (2018). doi:10.1016/j.autcon.2018.07.007.
[46] J. Chen, H. Chen, X. Luo, Collecting building occupancy data of high resolution based WiFi and BLE network, Autom. Constr. 102 (2019).
[47] A. Wagner, W. O'Brien, B. Dong, Exploring occupant behavior in buildings : methods and challenges, Springer International Publishing, 2017. doi:10.1007/978-3-319-61464-9.
[48] Z. Wang, T. Hong, M.A. Piette, Data fusion in predicting internal heat gains for office buildings through a deep learning approach, Appl. Energy. 240 (2019) 386–398. doi:10.1016/J.APENERGY.2019.02.066.
[49] Z. Chen, M.K. Masood, Y.C. Soh, A fusion framework for occupancy estimation in office buildings based on environmental sensor data, Energy Build. 133 (2016) 790–798. doi:10.1016/J.ENBUILD.2016.10.030.
[50] T. Hong, S.H. Lee, Integrating physics-based models with sensor data: An inverse modeling approach, Build. Environ. 154 (2019) 23–31. doi:10.1016/J.BUILDENV.2019.03.006.
[51] A. Szczurek, M. Maciejewska, M. Teuerle, A. Wyłomańska, Method to characterize collective impact of factors on indoor air, Phys. A Stat. Mech. Its Appl. 420 (2015) 190–199. doi:10.1016/J.PHYSA.2014.10.094.